JOURNAL OF CLINICAL MICROBIOLOGY, Sept. 2002, p. 3319–3325 Vol. 40, No. 9
0095-1137/02/$04.00+0 DOI: 10.1128/JCM.40.9.3319–3325.2002

Rational Design of DNA Sequence-Based Strategies for Subtyping *Listeria monocytogenes*

Steven Cai,[1] Dirce Yorika Kabuki,[1]† Arnaldo Yoshiteru Kuaye,[1]† Theresa Gina Cargioli,[1] Michael S. Chung,[1] Rasmus Nielsen,[2] and Martin Wiedmann[1]*

Department of Food Science[1] and Department of Biometry,[2] Cornell University, Ithaca, New York

Received 23 April 2002/Returned for modification 29 May 2002/Accepted 1 July 2002

The ability to differentiate bacteria beyond the species level is essential for identifying and tracking infectious disease outbreaks and to improve our knowledge of the population genetics, epidemiology, and ecology of bacterial pathogens. Commonly used subtyping methods, such as serotyping, phage typing, ribotyping, and pulsed-field gel electrophoresis, can yield ambiguous results that are difficult to standardize and share among laboratories. DNA sequence-based subtyping strategies can reduce interpretation ambiguity. We report the development of a rational approach for designing sequence-based subtyping methods. *Listeria monocytogenes* was selected as the model organism for testing the efficacy of this approach. Two housekeeping genes (*recA* and *prs*), one stress response gene (*sigB*), two virulence genes (*actA* and *inlA*), and two intergenic regions (*hly-mpl* and *plcA-hly*) were sequenced for 15 *L. monocytogenes* isolates. Isolates were chosen from a representative collection of more than 1,000 *L. monocytogenes* isolates to reflect the genetic diversity of this species. DNA sequences were aligned, and sliding window analyses were performed for each gene to define 600-bp-long regions that were (i) most polymorphic (using ProSeq) or (ii) most discriminatory (using a new algorithm implemented in WINDOWMIN). Complete gene sequences for *actA* (1,929 bp) and *inlA* (2,235 bp) provided the highest discrimination (identifying 15 and 14 allelic types, respectively). WINDOWMIN allowed identification of 600-bp regions within these genes that provided similar discriminatory power (yielding 15 and 13 allelic types, respectively). The most discriminatory 600-bp fragments identified in the housekeeping and stress response genes differentiated the isolates into 8 to 10 subtypes; intergenic region sequences yielded 8 and 12 allelic types based on 335- and 242-bp sequences for *hly-mpl* and *plcA-hly*, respectively. Regions identified as most polymorphic were not necessarily most discriminatory; therefore, application of the WINDOWMIN algorithm provided a powerful tool for determining the best target regions for DNA sequence-based subtyping. Our specific results also show that inclusion of virulence gene target sequences in a DNA sequence-based subtyping scheme for *L. monocytogenes* is necessary to achieve maximum subtype differentiation.

Listeria monocytogenes is a food-borne pathogen that causes approximately 2,500 cases of human illness and 500 deaths annually in the United States (21). Bacterial subtyping methods have improved our ability to detect and track human listeriosis outbreaks and have also provided tools for tracking sources of *L. monocytogenes* contamination throughout food systems. Application of subtyping methods also provides insight into the population genetics, epidemiology, ecology, and evolution of *L. monocytogenes*. A variety of conventional, phenotypic, and DNA-based subtyping methods have been described for differentiation of *L. monocytogenes* beyond the species and subspecies levels (14). While phenotype-based methods have been used for many years to subtype *L. monocytogenes* and other food-borne pathogens, DNA-based subtyping methods are generally more discriminatory and amenable to interlaboratory standardization and are thus increasingly replacing phenotype-based subtyping methods (29).

Commonly used phenotype-based subtyping methods for *L. monocytogenes* and other food-borne pathogens include serotyping, phage typing, and multilocus enzyme electrophoresis (MLEE) (14, 24). DNA-based subtyping methods include PCR-based approaches (e.g., random amplified polymorphic DNA and amplified fragment length polymorphism), ribotyping, and pulsed-field gel electrophoresis (2, 13, 16). These DNA-based methods define bacterial subtypes by using either PCR amplification or restriction digestion of bacterial DNA to generate DNA fragment banding patterns. While many of these methods have proven effective for differentiating *L. monocytogenes* subtypes, DNA fragment size-based subtyping methods have significant drawbacks. For example, despite the existence of software packages for data normalization and analyses (12), these subtyping methods are often difficult to standardize. As a consequence, the ease of exchanging and comparing subtype data among laboratories can be severely limited. While DNA fragment size-based subtyping methods have been used for cluster analyses, they generally do not provide information amenable to the inference of primary genetic characteristics (i.e., nucleotide sequences) for evolutionary analyses. As long-term studies on the epidemiology, ecology, and evolution of bacterial pathogens require subtyping data that can be used to infer and quantify the genetic relatedness of isolates, DNA fragment size-based subtyping methods have limited utility for these applications.

DNA sequencing-based methods are being developed and increasingly used for subtyping and characterizing bacterial

* Corresponding author. Mailing address: Cornell University, Department of Food Science, 412B Stocking Hall, Ithaca, NY 14853. Phone: (607) 254-2838. Fax: (607) 254-4868. E-mail: mw16@cornell.edu.

† Present address: Faculdade de Engenharia de Alimentos, Universidade Estadual de Campinas, 13083-970 Campinas, SP, Brazil.

TABLE 1. *L. monocytogenes* isolates used in this study

Isolate no.	Previous isolate no.	Source[a]	Origin	Serotype	*hly* type[c]	*actA* type[c]	*Eco*RI ribotype[c]	Lineage[c]
FSL J1-038	HO413	CDC	Human	1/2b	1	4	DUP-1042	I
FSL J2-039	NA[b]	CUVDL	Turkey	4b	1	3	DUP-1042	I
FSL J2-045	NA	CUVDL	Sheep	4b	1	4	DUP-1042	I
FSL J1-051	HO436	CDC	Human	4b	1	4	DUP-1052	I
FSL J2-064	NA	CUVDL	Cow	1/2b	1	3	DUP-1052	I
FSL J2-003	NA	CUVDL	Cow	1/2a	2	3	DUP-1030	II
FSL J2-017	NA	CUVDL	Cow	1/2a	2	4	DUP-1030	II
FSL J1-022	DD942, NCTC 4885	Qualicon, Inc.	Human	1/2c	2	4	DUP-1030	II
FSL J1-047	HO478	CDC	Human	1/2c	2	4	DUP-1039	II
FSL C1-117	2007202	CTDOH	Human	1/2a	2	4	DUP-1039	II
FSL J2-068	NA	CUVDL	Horse	4c	1b	4	DUP-1059	III
FSL X1-002	L99	S. Notermans	Food	4a	4a	3	DUP-1059	III
FSL W1-110	DD 3823	Qualicon, Inc.	NA	4c	4a	3	DUP-1055	III
FSL W1-111	DD 6821	Qualicon, Inc.	NA	4c	4b	3	DUP-1032	III
FSL J1-158	NA	CUFSL	Goat	4b	4b	3	DUP-10142	III

[a] CDC, Centers for Disease Control and Prevention; CUVDL, Cornell University Veterinary Diagnostic Laboratory; CTDOH, Connecticut Department of Health; CUFSL, Cornell University Food Safety Laboratory.
[b] NA, information not available.
[c] Determinations of *hly* types, *actA* types, *Eco*RI ribotypes, and lineages were performed as described in references 16 and 30.

isolates. In these methods, complete or partial nucleotide sequences are determined for one or more bacterial genes or chromosomal regions, thus providing unambiguous and discrete data. Sequencing can target a single gene (single locus approach) or multiple genes. The advantages of sequencing methods over DNA fragment size-based typing methods include their ability to generate unambiguous data that are portable through web-based databases and that can be used for phylogenetic analyses (3, 9). While a variety of DNA sequence-based subtyping strategies targeting virulence genes, housekeeping genes, or other chromosomal genes and regions are feasible, multilocus sequence typing (MLST), which is an extension of MLEE, represents a widely used strategy (26).

MLEE differentiates bacterial strains by detecting variations in the patterns of the electrophoretic mobilities of various constitutive enzymes. Cell extracts containing soluble enzymes are separated by size in nondenaturing starch gels, and enzyme activities are determined in the gels through application of color-generating substrates (14). While MLEE has been used to study the population genetics of many bacterial pathogens, including *L. monocytogenes* (23), this method is difficult to standardize among laboratories.

MLST directly determines the allelic variation of multiple housekeeping genes by using DNA sequencing instead of indirectly characterizing these alleles via measurement of the electrophoretic mobilities of the gene products through MLEE (7). MLST approaches have been developed for several organisms, including group A streptococci, *Staphylococcus aureus*, *Neisseria meningitidis*, and *Campylobacter jejuni* (4–6). MLST traditionally targets multiple loci that have slowly diversified from each other through accumulation of neutral or near neutral changes, thus providing reliable differentiation without the potentially confounding effects of positive selection that may particularly occur in certain categories of genes, such as bacterial virulence or surface genes (20, 26). On the other hand, direct sequencing of virulence genes (9) or intergenic regions may provide more-sensitive discrimination for cluster analyses and short-term epidemiological questions.

While some studies have explored the suitability of sequencing single genes to differentiate *L. monocytogenes* strains (10, 28), the discriminatory power of DNA sequencing strategies that target multiple distinct genes or regions has not yet been reported. Thus, we selected a well-characterized set of *L. monocytogenes* isolates to determine the suitability of DNA sequencing of housekeeping and stress response genes (*sigB*, *prs*, and *recA*), two virulence genes (*actA* and *inlA*), and two intergenic regions (*plcA-hly* and *hly-mpl*) to differentiate *L. monocytogenes* subtypes. The complete sequence information for these genes and regions was also used to define the most-discriminatory 600-bp fragments within these genes that could be used for rapid sequencing-based subtyping. This study also provides a general outline for a rational approach to the selection of target genes for DNA sequence-based subtyping of bacterial pathogens.

MATERIALS AND METHODS

Bacterial strains and lysate preparation. A test set of 15 isolates, which includes 5 isolates from each of the three *L. monocytogenes* genetic lineages (30) (Table 1), was selected from a collection of more than 1,000 isolates from foods, environmental samples, and human and animal listeriosis cases. Within lineages I and II, isolates were selected to represent sets of two or three closely related strains to allow us to best determine the ability of different target genes to

TABLE 2. Relevant characteristics of target genes and regions used for DNA sequencing

Gene or region	Functional property	Location on EGD chromosome (kb)[a]
actA	Virulence gene (protein involved in actin tail formation)	209.5
inlA	Virulence gene (internalin, involved in host cell invasion)	454.7
prs	Phosphoribosyl synthetase	202.6
sigB	Stress responsive alternative sigma factor B	930.7
recA	Recombinase A	1,425.5
plcA-hly	Intergenic region	206
hly-mpl	Intergenic region	208

[a] Locations are based on the *L. monocytogenes* EGD genome (http://genolist.pasteur.fr/ListiList/index.html).

TABLE 3. PCR primers

Primer name	Primer target and direction	Sequence (5′→3′)
LmsigB-15	*sigB,* forward	AATATATTAATGAAAAGCAGGTGGAG
LmsigB-16	*sigB,* reverse	ATAAATTATTTGATTCAACTGCCTT
Prs-F	*prs,* forward	GCGCGAAACAAGACAACAAAC
LMVI-prfA	*prs,* reverse	GC(T/A)GTTAG(C/T)AGAATT(C/A)TTTC
ActA-F_{176}	*actA,* forward	TAGCGTATCACGAGGAGG
ActA-R_{2148}	*actA,* reverse	TTTTGAATTTCATATCATTCACC
ActA-CF	*actA,* forward, lineage III	TTTATGCGTGCGATGATGGTAGTT
ActA-DR	*actA,* reverse, lineage III	CTTTTTCGTCTTCTGCACTTTTAG
2plcB-ecoR	*actA,* reverse, FSL W1-110 and FSL X1-002	GGAATTCTTTATGTGGTAATTTGCTGTC
recA-F23	*recA,* forward	GTTCGCTTTTTGCTTGCTTCTC
recA-R1368	*recA,* reverse	AAAAACTTCCCCCTTGTCTCC
$inlAF_1$	*inlA,* forward	CAGGCAGCTACAATTACACA
$inlAR_1$	*inlA,* reverse	ATATAGTCCGAAAACCACATCT
hlympl-F	*hly-mpl,* forward	TCTCCATCTGGGGCACTACAC
hlympl-R	*hly-mpl,* reverse	TTACCATGATGATACAAATA
plcAhly-F	*plcA-hly* forward	CCGCCTAATGGGAAAGTAAA
plcAhly-R	*plcA-hly* reverse	AGGCGGAGATGCTGGTG

differentiate closely related genotypes. Lineage III isolates were chosen to represent one set of closely related strains and three diverse strains, as strains classified into lineage III appear more genetically diverse than strains in the other two lineages. Bacteria were grown on brain heart infusion (Difco, Sparks, Md.) agar and in brain heart infusion broth at 37°C. All stock cultures were stored at −80°C in 15% glycerol. Bacterial cell lysates for use in PCRs were prepared with lysozyme and proteinase K as previously described (11).

Genes selected for sequence typing. Two housekeeping genes (*prs* and *recA*), one stress response gene (*sigB*), two virulence genes (*actA* and *inlA*), and two intergenic regions in the *prfA* virulence gene cluster were sequenced (Table 2). The relative location of these genes on the *L. monocytogenes* chromosome is also shown in Table 2. The intergenic regions *plcA-hly* and *hly-mpl* were chosen because they constitute the largest intergenic regions found within the *prfA* virulence gene island.

PCR amplification conditions. All PCR amplifications were performed with *Thermus aquaticus* DNA polymerase (Perkin Elmer-Applied Biosystems, Foster City, Calif.), $MgCl_2$ at a final concentration of 1.5 mM, 1× PCR buffer, and deoxynucleoside triphosphates at a final concentration of 50 μM. PCR primers and conditions are summarized in Tables 3 and 4, respectively. PCR conditions were not optimized to achieve a single set of PCR conditions for all amplifications.

Purification and quantification of PCR products. For sequencing, PCR products were purified by using the Qiaquick PCR purification kit (Qiagen, Inc., Valencia, Calif.). For eight *L. monocytogenes* isolates, nonspecific amplification products and/or primer dimers >60 bp were observed with the *sigB* PCR primers. For these PCRs, we used the Qiaquick gel extraction kit (Qiagen) to isolate the desired PCR product following electrophoresis in a 1.5% low-melting-point agarose gel and excision of the appropriate band from the ethidium bromide-stained gel. DNA concentrations of purified PCR products were estimated by comparison of amplicon band intensity with that of a DNA marker of known size and concentration (pGEM marker; Promega, Madison, Wis.) with LabImage software (Kapelan, Halle, Germany).

DNA sequencing and analyses. DNA sequencing was performed at Cornell University's Bioresource Center with an ABI 3700 DNA Sequencer. With the exception of *actA* and *inlA*, PCR primers were also used for DNA sequencing. Following initial sequencing with PCR primers, primer walking with primers designed to match strain-specific internal sequences was used to complete the sequencing of *actA* and *inlA*. Nucleic acid sequences were proofread and aligned with Seqman (DNAStar) and MegAlign (Lasergene). Cluster analyses were conducted using MEGA, version 2.1 (19), and the unweighted pair group method with arithmetic mean (UPGMA) (number of nucleotide differences) model. The resulting clustering data were assessed together with the alignments to assign allelic types for each region or gene. Two sequences were assigned different allele numbers if the sequences differed by at least 1 nucleotide (nt). Sliding window analyses to determine the number of DNA polymorphisms per 600-bp window for each gene were performed with ProSeq software (http://helios.bto.ed.ac.uk/evolgen/filatov/proseq.html).

A new algorithm was developed and implemented for identifying the most-discriminatory 600-bp region within each gene from a set of aligned sequences (WINDOWMIN). In this algorithm, the most-discriminatory region for a sample of size n is defined as the region that maximizes the number of different DNA sequences in a sample. WINDOWMIN allows the user to define the criteria that will classify two DNA sequences as different. For example, if sequencing errors are common, two DNA sequences may be defined as different from each other if they differ at more than 2 nt positions. If sequencing errors can be ruled out, it may be more reasonable to define two DNA sequences as being different if they differ in at least 1 nt position. For this study, two DNA sequences were defined as different if they differed by at least 1 nt. In WINDOWMIN, the number of different DNA sequences distinguishable by at least k nucleotides in a window starting at position i of the sequences is defined as $n_{k,i}$. A window starting at position j in the sequence is most discriminatory if $_{nk,j} = \max_i(_{nk,i})$ for a particular value of k. If many possible windows fulfill this criteria, the procedure can be iterated to find the set of windows fulfilling $n_{k+1,j} = \max_i(n_{k+1,i})$ among the sequences fulfilling $n_{k,j} = \max_i(n_{k,i})$. Window size must also be defined; for the analyses performed here, the window size was set to 600 bp. As a practical approach, we iterated the algorithm 10 times and then chose the window with the most segregating sites among the set of remaining windows. WINDOWMIN can be obtained by download at http://www.foodscience.cornell.edu/wiedmann/programs.html.

Nucleotide sequence accession number. The DNA sequences determined in this study have been deposited in the GenBank database and given accession numbers AF497139 through AF497243.

TABLE 4. PCR conditions

Gene or regions	Initial denaturation	Amplification (no. of cycles)[a]	Final hold
actA	94°C, 2 min 30 s	94°C, 3 min; 53°C, 1 min; 72°C, 2 min (40)	72°C, 5 min
inlA	94°C, 2 min 30 s	94°C, 3 min; 66°C, 1 min; 72°C, 2.5 min (35)	72°C, 5 min
sigB	94°C, 5 min	94°C, 30 s; TD 54–44°C, 30 s; 72°C, 1 min (40)	72°C, 7 min
prs	94°C, 5 min	94°C, 30 s; TD 60–50°C, 30 s; 72°C, 1 min (40)	72°C, 7 min
recA	94°C, 2 min 30 s	94°C, 30 s; 50°C, 30 s; 72°C, 1 min (40)	72°C, 7 min
hly-mpl and *plcA-hly*	94°C, 2 min 30 s	94°C, 30 s; 54°C, 30 s; 72°C, 1 min (40)	72°C, 7 min

[a] TD, touchdown PCR. For *sigB* and *prs,* a touchdown PCR protocol was used with annealing temperatures decreasing at a rate of 0.5°C/cycle from an annealing temperature T for the first 20 cycles followed by another 20 cycles with the annealing temperature set at $T - 10$°C.

TABLE 5. Summary of allelic subtypes and polymorphisms in *actA, inlA, prs, sigB, recA, hly-mpl,* and *plcA-hly*

Gene or intergenic region	Full sequence			Optimal window with PROSEQ (highest no. of polymorphisms)			Optimal window with WINDOWMIN (best descrimination)		
	Length (bp)	No. of allelic subtypes[a]	No. of polymorphic sites[b]	Location (nt)[c]	No. of allelic subtypes[a]	No. of polymorphic sites[b]	Location (nt)[c]	No. of allelic subtypes[a]	No. of polymorphic sites[b]
actA	1,929	15 (14)	276 (14.3)	1,330–1,929	13 (9)	104 (17.3)	316–915	15 (14)	86 (14.3)
inlA	2,235	14 (14)	186 (8.3)	964–1,563	13 (11)	66 (11.0)	1,090–1,690	13 (13)	65 (10.8)
sigB	780	10 (10)	79 (10.1)	13–612	9 (8)	74 (12.3)	150–749	10 (10)	58 (9.7)
recA	1,047	8 (7)	107 (10.2)	264–863	8 (7)	67 (11.1)	238–837	8 (7)	66 (11.0)
prs	957	10 (7)	47 (4.9)	316–915	10 (7)	42 (7.0)	313–912	10 (7)	41 (6.8)
hly-mpl	335	8 (7)	28 (8.4)	NA[d]	NA	NA	NA	NA	NA
plcA-hly	242	12 (8)	26 (10.7)	NA	NA	NA	NA	NA	NA

[a] Total number of allelic subtypes distinguished within the specified region by a difference of 1 nt. Numbers in parentheses denote the total number of alleles distinguished by using the criteria of 2 nt differences.
[b] Number of polymorphic sites within the specified region. Values in parentheses are the percentages of polymorphic sites out of all nucleotides for that specific region.
[c] Locations are nucleotide (nt) positions (nucleotide 1 is the first nucleotide of the coding sequence).
[d] NA, not available.

RESULTS

Sequencing results. The complete open reading frames (ORFs) of two virulence genes (*actA* and *inlA*), three housekeeping and stress response genes (*prs*, *recA*, and *sigB*), and two intergenic regions were sequenced for 15 *L. monocytogenes* isolates (Tables 1 and 2). Overall, approximately 7,500 bp of genomic DNA were sequenced for each strain. DNA sequences were deposited into GenBank and are also available through the Cornell PathogenTracker database (http://pathogentracker.net).

Analyses of DNA polymorphisms and allelic types. DNA sequences for each of the seven selected genes and intergenic sequences were aligned and evaluated. The size of these seven target sequences varied from 242 bp (*plcA-hly* intergenic sequence) to 2,235 bp (*inlA*) (Table 5). Based on the total number of polymorphic sites in each sequence (Table 5), *actA* displayed the highest sequence variability (14.3% of the nucleotides were polymorphic) while *prs* displayed the lowest level of overall sequence variability (4.9%). Only one target sequence (*actA*) allowed discrimination of all 15 strains characterized. Two strains (FSL J1-022 and FSL J1-047) differed by only a single nucleotide in *actA*. The other target genes and regions differentiated between 8 and 14 allelic types for each gene (Table 5). Two sequences were assigned different allelic types (e.g., 1 to 15 for *actA*) if the sequences differed by at least 1 nt; the allelic types for all strains are summarized in Table 6. Table 5 also shows how many allelic subtypes were defined based on at least a single nucleotide polymorphism and how many subtypes could be defined based on a cutoff of at least a 2-nt difference. For example, for the full-length *prs* sequence 10 allelic subtypes were defined based on at least a 1-nt difference and 7 allelic subtypes were defined based on at least a 2-nt difference. Thus, three of the *prs* allelic types were defined based on only a 1-nt difference. In addition, allelic information for the housekeeping genes *prs* and *recA* and the stress response gene *sigB* was used to assign MLST types (12 MLST types, A through L) (Table 6) based on the allelic types for all three genes.

Insertion or deletion polymorphisms in *actA*, *inlA*, and *hly-mpl*. In addition to single nucleotide polymorphisms, sequence alignments revealed insertion-deletions in *actA*, *inlA*, and *hly-mpl*. Nucleotide sequence alignment for *actA* revealed a deletion of 105 bp in strains FSL W1-110, FSL W1-111, FSL J2-039, FSL J1-158, FSL J2-003, FSL J2-064, and FSL X1-002. This finding is consistent with a previous study that reported deletion of a proline-rich repeat region among isolates classified into each of the three genotypic lineages of *L. monocytogenes* (30). Moreover, isolates FSL W1-111 and FSL J1-158 showed one insertion or deletion at *actA* nt 323 (a 9-bp insertion compared to the other isolates). The *inlA* sequence alignment revealed an insertion or deletion at nt 2052 in three strains (a 9-bp deletion in FSL W1-110, FSL J2-068, and FSL J2-003) and another 3-bp insertion or deletion at nt 2227 in strains FSL W1-111 and FSL J1-158 (a 3-bp insertion in these two isolates). Furthermore, isolate FSL J1-158 had a single base pair deletion at nt 2219, which results in a frameshift mutation with 7 missense amino acid residues at the 3′ end of the *inlA* sequence.

Sequence analyses of the *hly-mpl* intergenic region also revealed lineage-specific insertions or deletions. The five lineage I strains show two additional nucleotides at position 49. All lineage II strains have an additional adenine at nt 52. Insertion

TABLE 6. Allelic profiles of the virulence genes (*actA* and *inlA*), intergenic regions (*hly-mpl* and *plcA-hly*), and housekeeping genes (*sigB, prs,* and *recA*)

Strain	Allelic designation based on:							MLST type based on *sigB, prs,* and *recA*
	actA	*inlA*	*hly-mpl*	*plcA-hly*	*sigB*	*prs*	*recA*	
FSL J1-038	1	1	1	1	1	1	1	A
FSL J2-039	2	2	1	2	1	2	2	B
FSL J2-045	3	3	1	3	2	1	2	C
FSL J1-051	4	4	1	3	2	3	2	D
FSL J2-064	5	5	1	2	3	4	2	E
FSL J2-003	6	6	2	4	4	1	3	F
FSL J2-017	7	7	3	5	5	5	3	G
FSL J1-022	8	8	2	6	5	5	3	G
FSL J1-047	9	8	2	6	5	5	3	G
FSL C1-117	10	9	2	7	5	5	3	G
FSL J2-068	11	10	4	8	6	6	4	H
FSL X1-002	12	11	5	9	7	7	5	I
FSL W1-110	13	12	6	10	8	8	6	J
FSL W1-111	14	13	7	11	9	9	7	K
FSL J1-158	15	14	8	12	10	10	8	L

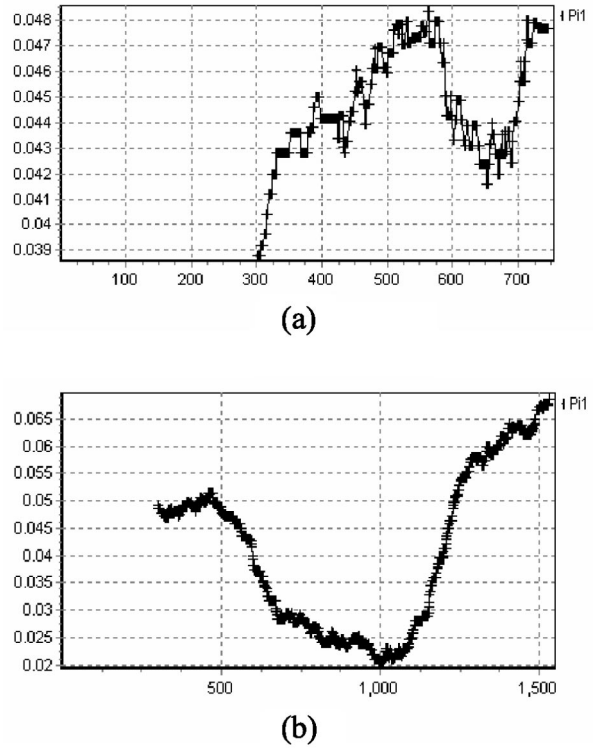

FIG. 1. Graphical representation of polymorphisms within 600-bp sliding windows in the genes *recA* (a) and *actA* (b). Pi denotes the average number of nucleotide differences per site between two sequences, or nucleotide diversity (22). Pi was calculated with the computer program ProSeq (see Materials and Methods).

or deletion polymorphisms at the nt 121 to 129 region also show distinct patterns that differentiate strains in lineages I, II, and III; lineage II isolates have a 1-bp deletion compared to lineage I isolates, whereas lineage III isolates show 9- and 8-bp deletions compared the lineage I and II isolates, respectively.

Definition of most-discriminatory gene fragments. Since complete sequencing of large DNA fragments (e.g., the 1,929-bp *actA* ORF) is not practical, in terms of cost and time, for large-scale subtyping, we utilized two different approaches to define smaller (600 bp) gene fragments that will provide optimal strain differentiation, ideally with the same discriminatory power achieved by analysis of the complete ORF sequences. The software ProSeq (version 2.8) was used initially in a sliding window analysis to define the 600-bp region within each gene (*actA*, *inlA*, *recA*, *sigB*, and *prs*) that showed the highest number of nucleotide polymorphisms. ProSeq calculated the value π, which is the average number of nucleotide differences per site between two sequences (also termed nucleotide diversity) (22), for each gene from the nucleotide alignment of the 15 strains. The region with the maximum nucleotide diversity is shown as the highest peak on a plot of π determined at sliding windows of 600 bp (Fig. 1). The locations of the 600-bp regions with the maximum nucleotide diversities for each gene are shown in Table 5. For *actA*, *inlA*, and *sigB*, the 600-bp window with the maximum nucleotide diversity did not allow the same level of allelic subtype differentiation as was achieved with the full ORF sequences (Table 5).

DNA sequence alignments were also used to perform sliding window analyses for each gene to define 600-bp sections that were most discriminatory for differentiating the aligned sequences into allelic types. An algorithm (WINDOWMIN) was implemented to determine the most-discriminatory 600-bp section. For three (*actA*, *inlA*, and *sigB*) of the five gene alignments, the WINDOWMIN sliding window analysis defined 600-bp regions that showed superior discrimination of allelic subtypes compared to the regions with the highest nucleotide diversities defined by the ProSeq sliding window analysis (Table 5). Specifically, the 600-bp windows defined by WINDOWMIN allowed subtype discrimination equal to that achieved with the complete gene sequences for four of the genes sequenced (*actA*, *recA*, *prs*, and *sigB*), which ranged in length from 780 to 1,929 bp. WINDOWMIN was not able to define a 600-bp region that matched the allelic subtype discrimination of the full ORF for *inlA*, the largest gene sequenced (2,235 bp). The full *inlA* ORF allowed differentiation of 14 allelic subtypes, whereas the most-discriminatory 600-bp sequence identified by WINDOWMIN allowed differentiation of 13 allelic subtypes.

DISCUSSION

We used *L. monocytogenes* as a model organism to develop and apply a rational approach for designing a DNA sequence-based subtyping scheme. Two housekeeping genes, one stress response gene, two virulence genes, and two intergenic regions were sequenced for 15 *L. monocytogenes* isolates selected from a collection of more than 1,000 isolates. The DNA sequence data obtained (approximately 7,500 bp/isolate) were used to define both the most-polymorphic and the most-discriminatory 600-bp regions within each gene. This model for designing discriminatory DNA sequence-based subtyping strategies also should be applicable to other target organisms. The development of a new algorithm designed to determine the most discriminatory section within a DNA sequence alignment and the implementation of this algorithm in a new program (WINDOWMIN) will further facilitate the definition of target genes and gene fragments for subtyping applications.

Target genes for DNA sequence-based subtyping of *L. monocytogenes*. Housekeeping genes are commonly used for conventional MLST as these genes are thought to diversify by neutral or near neutral nucleotide changes due to the vital roles of these gene products in contributing to an organism's survival (7, 9). While sequence analysis of housekeeping genes has been a valuable tool for studying the population genetics of bacterial pathogens, DNA sequencing of more-rapidly evolving genes may allow more-sensitive subtype discrimination. We thus selected the housekeeping genes *recA* and *prs*, the stress response gene *sigB*, the virulence genes *actA* and *inlA*, and the intergenic regions *hly-mpl* and *plcA-hly* to determine the relative abilities of these target genes to differentiate closely related *L. monocytogenes* strains. While *prs* is located close to *actA* on the *L. monocytogenes* chromosome, *sigB*, *recA*, and the virulence genes were chosen to represent distinct chromosomal locations. *prs* was chosen as a target gene that would allow comparison between the discriminatory power of virulence genes and housekeeping genes located in close proximity. *actA* and *inlA* were specifically selected as target virulence genes since they are located in two different *L. monocytogenes* virulence gene islands (18). The complete gene sequences for

actA and *inlA* provided the highest discrimination among the 15 isolates (15 and 14 allelic types, respectively). In addition, preliminary analyses of our data showed that the *actA* virulence gene indeed may be under positive selection ($d_N/d_S > 1.0$; d_N, rate of nonsynonymous substitutions; d_S, rate of synonymous substitutions) (unpublished data). The complete gene sequences for the housekeeping and stress response genes provided discrimination into 8 to 10 subtypes, and the sequences for the intergenic regions provided 8 and 12 allelic types based on 335- and 242-bp sequences for *hly-mpl* and *plcA-hly*, respectively. When sequence information for the two housekeeping genes and the stress response gene was used to define MLST types, a total of 12 types (A to L) (Table 6) could be differentiated. These results indicate that sequencing of the virulence genes *actA* and *inlA* provides the most-discriminatory DNA sequence-based subtyping for *L. monocytogenes*. While limited sequencing and PCR-restriction fragment length polymorphism analysis of other *L. monocytogenes* virulence genes has been performed by various groups (10, 27, 28, 30), the results reported here provide the first comparative evaluation of different target genes for subtyping *L. monocytogenes*.

Our results support previous observations that DNA sequencing of virulence or surface protein-encoding genes that may have been subjected to positive selection pressures can allow more-sensitive strain discrimination than sequencing of housekeeping genes can. To illustrate, Enright et al. (9) showed that *emm* sequences differentiated more subtypes than single housekeeping gene sequences in *Streptococcus pyogenes*. However, interpretation of subtyping results based on highly variable sequences of virulence genes (such as *actA*) or surface proteins may be misleading, particularly if the data are used to probe the population genetics or long-term phylogenetic patterns of bacterial pathogens such as *L. monocytogenes*. Specifically, high rates of evolution and recombination among these genes may not reflect the true phylogenetic relationships among isolates (26). To overcome this obstacle, a subtyping scheme that includes sequencing of selected virulence genes in combination with sequencing of housekeeping and/or stress response genes and of regions with little or no selective pressure (such as intergenic regions) may provide the most appropriate approach for subtyping *L. monocytogenes* and other bacterial pathogens. Initial analyses for recombination within the two housekeeping genes and one stress response gene sequenced here indeed showed that these genes show no (*sigB* and *recA*) or weak (*prs*) indication for recombination (unpublished data). The sequencing of additional housekeeping genes as previously described for MLST approaches (4–6) for other bacteria may further improve the ability to study the phylogeny of *L. monocytogenes*. The inclusion of both positively selected genes, such as *actA*, and neutrally selected intergenic regions (such as the more-discriminatory *plcA-hly* region) allows maximization of the discriminatory power of a typing scheme though. Maximum discrimination is particularly important for bacterial pathogens such as *L. monocytogenes*, for which rapid and standardized cluster detection through molecular subtyping represents a critical public health need (29). The use of virulence gene targets in DNA sequence-based subtyping strategies also creates the opportunity to use pathogen-specific PCR primers to develop integrated PCR-based detection and subtyping strategies that do not require a culturing step (5).

Rational design of high-throughput DNA sequence-based subtyping schemes. Sequencing of complete virulence, housekeeping, and stress response genes (with ORF lengths between 780 and 2,235 bp) (Table 5) does not provide a suitable approach for high-throughput subtyping. High-throughput sequencing for subtyping purposes generally targets gene fragments between 450 and 600 bp in length, since these fragment sizes can easily be amplified and sequenced with a single set of primers or one set of sequencing primers nested inside the PCR primers (7, 20). In the past, target regions have been selected without prior identification of specific desirable characteristics, such as an optimum number of polymorphic and discriminatory nucleotide sites. To identify target regions through a rational strategy, the complete ORFs for the five genes sequenced in this study were aligned to define the 600-bp section(s) for each gene that was (i) most polymorphic (with ProSeq) or (ii) most discriminatory (with WINDOWMIN). Not surprisingly, most genes, and particularly the larger virulence genes *actA* and *inlA*, displayed considerable differences in the numbers of polymorphic residues found in different regions. Thus, the allelic discrimination achieved with different 600-bp regions within a given gene also differed considerably (Table 5). Interestingly, the 600-bp regions with the highest proportion of polymorphic residues did not necessarily provide the highest level of allelic discrimination. For three of the five genes sequenced (*actA*, *inlA*, and *sigB*), WINDOWMIN was able to define more discriminatory 600-bp regions than a program (ProSeq) that only determined the most-polymorphic 600-bp region within a gene. We conclude that our newly developed algorithm provides an improved rational approach for the selection of target regions for DNA sequence-based subtyping of bacterial pathogens and other microorganisms.

DNA sequence-based subtyping in *L. monocytogenes*. A variety of molecular subtyping approaches have been applied to *L. monocytogenes*, and the application of these techniques has allowed a better understanding of the biology, ecology, and epidemiology of *L. monocytogenes* and other bacterial pathogens (29). Studies using subtyping approaches also have suggested that *L. monocytogenes* subtypes may display heterogeneity in their potentials to cause disease in humans and animals (15, 16, 30). The data reported here provide the framework for the development and implementation of DNA sequence-based subtyping methods for *L. monocytogenes*. We have identified specific 600-bp regions that provide the most discriminatory targets for subtyping within different *L. monocytogenes* genes. The presence of insertion-deletions in some of these regions (e.g., at nt 852 to 1019 in *actA*) may complicate the interpretation of subtyping results for some isolates and may hamper design of PCR primers that allow reliable amplification of all *L. monocytogenes* subtypes. Thus, in addition to identifying the most discriminatory gene regions, the presence of insertion-deletions must also be carefully considered when selecting target regions for DNA sequence-based subtyping methods. Interestingly, our results also indicate that targeting specific insertion-deletions (e.g., in the *hly-mpl* intergenic region) by appropriate PCR assays may allow for sensitive differentiation of the three previously described *L. monocytogenes* lineages (17, 30). Phylogenetic analyses of housekeeping, stress response, and virulence gene sequences also confirmed that the 15 *L. monocytogenes* isolates tested fall into the previously

determined three distinct lineages (unpublished data) (30). The existence of these lineages has previously been confirmed by a variety of subtyping methods and thus appears to be evolutionarily relevant (29).

While current DNA fragment size-based subtyping methods (such as pulsed-field gel electrophoresis and ribotyping) may provide good subtype differentiation, data obtained by these methods typically cannot be used to determine the evolutionary relatedness of isolates. The implementation of DNA sequence-based subtyping approaches for routine characterization of human, animal, and food *L. monocytogenes* isolates will not only allow for sensitive and standardized subtyping for outbreak detection, but will also provide an opportunity for using subtyping data to probe the evolution of this food-borne pathogen and to track the spread of clonal groups (25, 29). DNA sequence-based subtyping methods will also provide standardized data that can easily be shared electronically and through the World Wide Web (26), concomitantly providing public health professionals and laboratories around the world with direct access to the information needed to identify and monitor emerging pathogenic bacteria. The resolution power of DNA sequence-based subtyping methods is unmatched by any other subtyping method. For example, while it is estimated that MLEE requires approximately 26 nt changes in order to determine a new electrophoretic type (1), one single nucleotide change at a targeted locus will result in a new subtype classification for DNA sequence-based subtyping (8). The continued development of new technologies for automated high-throughput sequencing and the availability of these technologies at a reasonably low cost will further facilitate widespread implementation of DNA sequence-based subtyping methods. Further application of the DNA sequencing-based subtyping approaches described here on large sets of epidemiologically well-defined isolates will also provide critical validation of the subtyping scheme proposed here.

ACKNOWLEDGMENTS

This work was supported in part by USDA National Research Initiative Award no. 99-35201-8074 (to M.W.) and National Institutes of Health Award no. R01GM63259 (to M.W.).

We thank Celine Nadon and Kathyrn Boor for help with this project and for critical review of the manuscript.

REFERENCES

1. **Boyd, E. F., K. Nelson, F. S. Wang, T. S. Whittam, and R. K. Selander.** 1994. Molecular genetic basis of allelic polymorphism in malate dehydrogenase (MDH) in natural populations of *Escherichia coli* and *Salmonella enterica*. Proc. Natl. Acad. Sci. USA **91:**1280–1284.
2. **Bruce, J. L., R. J. Hubner, E. M. Cole, C. I. McDowell, and J. A. Webster.** 1995. Sets of *Eco*RI fragments containing ribosomal RNA sequences are conserved among different strains of *Listeria monocytogenes*. Proc. Natl. Acad. Sci. USA **92:**5229–5233.
3. **Chan, M. S., M. C. J. Maiden, and B. G. Spratt.** 2001. Database-driven multi locus sequence typing (MLST) of bacterial pathogens. Bioinformatics **17:** 1077–1083.
4. **Dingle, K. E., F. M. Colles, D. R. A. Wareing, R. Ure, A. J. Fox, F. E. Bolton, H. J. Bootsma, J. L. Willems, R. Urwin, and M. C. J. Maiden.** 2001. Multilocus sequence typing system for *Campylobacter jejuni*. J. Clin. Microbiol. **39:**14–23.
5. **Enright, M., K. Knox, D. Griffiths, D. Crook, and B. Spratt.** 2000. Molecular typing of bacteria directly from cerebrospinal fluid. Eur. J. Clin. Microbiol. Infect. Dis. **19:**627–630.
6. **Enright, M. C., N. P. J. Day, C. E. Davies, S. J. Peacock, and B. G. Spratt.** 2000. Multilocus sequence typing for the characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. J. Clin. Microbiol. **38:**1008–1015.
7. **Enright, M. C., and B. G. Spratt.** 1999. Multilocus sequence typing. Trends Microbiol. **7:**482–487.
8. **Enright, M. C., and B. G. Spratt.** 1998. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. Microbiology **144:**3049–3060.
9. **Enright, M. C., B. G. Spratt, A. Kalia, J. H. Cross, and D. E. Bessen.** 2001. Multilocus sequence typing of *Streptococcus pyogenes* and the relationships between *emm* type and clone. Infect. Immun. **69:**2416–2427.
10. **Ericsson, H., H. Unnerstad, J. Mattsson, M. Danielsson-Tham, and W. Tham.** 2000. Molecular grouping of *Listeria monocytogenes* based on the sequence of the *inlB* gene. J. Med. Microbiol. **49:**73–80.
11. **Furrer, B., U. Candrian, C. Hoefelein, and J. Luethy.** 1991. Detection and identification of *Listeria monocytogenes* in cooked sausage products and in milk by in vitro amplification of haemolysin gene fragments. J. Appl. Bacteriol. **70:**372–379.
12. **Gerner-Smidt, P., L. M. Graves, S. Hunter, and B. Swaminathan.** 1998. Computerized analysis of restriction fragment length polymorphism patterns: comparative evaluation of two commercial software packages. J. Clin. Microbiol. **36:**1318–1323.
13. **Graves, L. M., and B. Swaminathan.** 2001. Pulsenet standardized protocol for subtyping *Listeria monocytogenes* by macrorestriction and pulsed-field gel electrophoresis. Int. J. Food Microbiol. **65:**55–62.
14. **Graves, L. M., B. Swaminathan, and S. B. Hunter.** 1999. Subtyping *Listeria monocytogenes*, p. 279–297. *In* E. Ryser and E. Marth (ed.), Listeria, listeriosis, and food safety. Marcel Dekker, New York, N.Y.
15. **Hof, H., and J. Rocourt.** 1992. Is any strain of *Listeria monocytogenes* detected in food a health risk? Int. J. Food Microbiol. **16:**173–182.
16. **Jeffers, G. T., J. L. Bruce, P. McDonough, J. Scarlett, K. J. Boor, and M. Wiedmann.** 2001. Comparative genetic characterization of *Listeria monocytogenes* isolates from human and animal listeriosis cases. Microbiology **147:** 1095–1104.
17. **Jinneman, K. C., and W. E. Hill.** 2001. Listeria monocytogenes lineage group classification by MAMA-PCR of the listeriolysin gene. Curr. Microbiol. **43:**129–133.
18. **Kreft, J., J.-A. Vazquez-Boland, E. Ng, and W. Goebel.** 1999. Virulence gene clusters and putative pathogenicity islands in listeriae, p. 219–232. *In* J. B. Kaper and J. Hacker (ed.), Pathogenicity islands and other mobile virulence elements. ASM Press, Washington, DC.
19. **Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei.** 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17:**1244–1245.
20. **Maiden, M. C. J., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. USA **95:**3140–3145.
21. **Mead, P., L. Slutsker, V. Dietz, L. F. McCaig, J. S. Bresee, C. Shapiro, P. Griffin, and R. V. Tauxe.** 1999. Food-related illness and death in the United States. Emerg. Infect. Dis. **5:**607–625.
22. **Nei, M.** 1987. Molecular evolutionary genetics. Columbia University Press, New York, N.Y.
23. **Piffaretti, J.-C., H. Kressebuch, M. Aeschenbacher, J. Bille, E. Bannerman, J. M. Musser, R. K. Seelander, and J. Rocourt.** 1989. Genetic characterization of clones of the bacterium *Listeria monocytogenes* causing epidemic disease. Proc. Natl. Acad. Sci. USA **86:**3818–3822.
24. **Schönberg, A., E. Bannerman, A. L. Courtieu, R. Kiss, J. McLaughlin, S. Shah, and D. Wilhelms.** 1996. Serotyping of 80 strains from the W. H. O. multicentre international typing study of *Listeria monocytogenes*. Int. J. Food Microbiol. **32:**279–287.
25. **Schuchat, A., B. Swaminathan, and C. Broome.** 1991. Epidemiology of human listeriosis. Clin. Microbiol. Rev. **4:**169–183.
26. **Spratt, B. G.** 1999. Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the internet. Curr. Opin. Microbiol. **2:**312–316.
27. **Vines, A., M. W. Reeves, S. Hunter, and B. Swaminathan.** 1992. Restriction fragment length polymorphism in four virulence-associated genes of *Listeria monocytogenes*. Res. Microbiol. **143:**281–294.
28. **Vines, A., and B. Swaminathan.** 1998. Identification and characterization of nucleotide sequence differences in three virulence-associated genes of *Listeria monocytogenes* strains representing clinically important serotypes. Curr. Microbiol. **36:**309–318.
29. **Wiedmann, M.** 2002. Molecular subtyping methods for *Listeria monocytogenes*. J. Assoc. Off. Anal. Chem. **85:**524–531.
30. **Wiedmann, M., J. L. Bruce, C. Keating, A. E. Johnson, P. L. McDonough, and C. A. Batt.** 1997. Ribotypes and virulence gene polymorphisms suggest three distinct *Listeria monocytogenes* lineages with differences in pathogenic potential. Infect. Immun. **65:**2707–2716.